# Estado del arte

Working with audio (tutorial) -> <https://www.youtube.com/watch?v=NYhkqXpFAlg&t=475s>

Modelo que detecta deepfake:

<https://huggingface.co/mo-thecreator/Deepfake-audio-detection>

<https://huggingface.co/spaces/Hiya-ai/loccus-audio-fake-detection>

Para generar voz (inglés)-> <https://colab.research.google.com/github/neonbjb/tortoise-tts/blob/main/tortoise_tts.ipynb>

Generar voz en español -> <https://huggingface.co/OuteAI/OuteTTS-1.0-0.6B>

<https://huggingface.co/coqui/XTTS-v2> (el que yo he usado)

Tutorial para feature extractor en audios -> <https://docs.pytorch.org/audio/stable/tutorials/audio_feature_extractions_tutorial.html#sphx-glr-tutorials-audio-feature-extractions-tutorial-py>

Synthetic voice detection -> <https://github.com/csun22/Synthetic-Voice-Detection-Vocoder-Artifacts?tab=readme-ov-file>

¿Cómo se detecta una voz real de una generada por IA?

Imagen en blanco y negro

El contenido generado por IA puede ser incorrecto.

Aquí vemos las ondas, que a simple vista, son similares, pero hay patrones acústicos que pueden diferenciar un audio real de uno generado.

1. Prosodia (ritmo, entonación y pausas)

* Voz real: Suele tener variaciones naturales en la entonación, duración de sílabas y pausas.
* Voz clonada: A menudo tiene un ritmo más mecánico o pausas ligeramente forzadas. Puede sonar demasiado perfecto o demasiado plano.

1. Microvariaciones del habla

En una voz real hay temblores, titubeos, aire, microentonaciones, ruidos involuntarios del entorno o del aparato fonador.

Las voces generadas tienden a ser más limpias y sin esas imperfecciones. Esto puede verse también en un análisis espectrograma (no solo waveform).

1. Textura del sonido

* Las voces reales tienen transiciones armónicas más suaves.
* Las sintéticas pueden tener bordes más definidos o ausencia de ruido de fondo natural.

1. Respuesta en frecuencia

Las voces sintéticas pueden tener menos contenido en ciertas bandas, especialmente en frecuencias agudas (ruido sibilante) o graves (sonidos guturales).

Para detectarlo automáticamente podríamos:

* Extraer MFCC (Mel-Frequency Cepstral Coefficients) o espectrogramas mel-scaled -> entrenar un modelo binario
* Usar modelos preentrenados: wav2vec2, whisper o audio-classification
* Añadir un sistema de detección basado en análisis estadístico de prosodia y duración fonética (yo creo que esto se va del scope).

Para el primer punto de MFCC -> Es una representación numérica del audio que intenta modelar cómo percibe el oído humano el sonido. Extrae características clave de cómo suena la voz, más allá del contenido, la firma acústica es diferente y se refleja en los coeficientes.

Lo de MFCC contiene transformada de Fourier (para descomponer audio en frecuencias), Mel Scale (una escala perceptiva basada en cómo los humanos interpretamos frecuencias) y Cepstrum (patrones derivados de la envolvente espectral).

Ahora en el workspace:

* Transcripciones.py: saca transcripciones de audios
* Jupyter: todas las pruebas
* Grabador\_audio.py: dices cuántos audios quieres y de que duración (se supone tienes ya el guion y lo vas leyendo)
* Huggin\_model.py: unas pruebas que hice para generar voz falsa pero no funcionó

Al parecer, se puede usar cualquier modelo de detección falsa sin necesidad que sea en español, pues la clave es que **el modelo no analiza el significado de las palabras, sino las "huellas" del audio.**

Los modelos que generan voces de IA, aunque son muy avanzados, a menudo dejan pequeños artefactos o patrones antinaturales en el sonido. Estas imperfecciones pueden ser:

* Un ligero zumbido metálico.
* Una cadencia o ritmo demasiado perfectos.
* Frecuencias de sonido que no son típicas de una voz humana.

Estos artefactos son, en gran medida, **independientes del idioma**. La física del sonido generado es lo que el modelo detecta, no la gramática del español.

<https://huggingface.co/MattyB95/AST-ASVspoof5-Synthetic-Voice-Detection?library=transformers> -> opción

Microsoft/wavlm-base-plus-sd -> opción

# Ideas de lo que podríamos hacer

* Sistema de confianza del modelo en tres niveles:
  + Confirmado es IA (si la confianza es muy alta)
  + Confirmado es humano
  + Incierto / sospechoso: La confianza para ambas etiquetas está en la zona media, es mejor esto que una clasificación incorrecta.
* Noisereduce: librería en Python, puede llegar a mejorar drásticamente la precisión del modelo. Es pasarle un filtro para reducir ruido de fondo y los artefactos de la llamada.
* Podríamos encontrar dos modelos de detección diferentes, y les vamos pasando el mismo fragmento de audio y comparando las respuestas. Si ambos coinciden, la confianza es muy alta, si no, se clasifica como incierto automáticamente.
* Para desplegar el modelo, podríamos hacer como una interfaz gráfica sencilla, que muestre el indicador de estado (humano, IA o incierto), puntuación de confianza y un botón para iniciar / detener el análisis. Incluso que, si se detecta que es IA, que la pantalla parpadee algo rojo, para que sea muy fácil de ver.
* Detección de palabras clave: sobre el texto transcrito, si se detectan palabras clave como “oferta”, “Banco”, “contraseña”, “bizum”, “datos personales”, “tarjeta de crédito”; Se aumenta el nivel de alerta

# Impacto legal

Dos normativas que rigen principalmente el proyecto:

1. Secreto de las comunicaciones ([Constitución Española, Art. 18.3](https://app.congreso.es/consti/constitucion/indice/titulos/articulos.jsp?ini=18&tipo=2)): Garantiza el secreto de las comunicaciones, especialmente las postales, telegráficas y telefónicas, salvo mediante resolución judicial. Esto cambia si eres un participante en la conversación, no es delito que uno de los interlocutores grabe la conversación.
2. Protección de datos personales (RGPD y LOPDGDD): La voz de una persona es un dato personal, ya que se puede identificar con ella. Aunque las grabaciones no se almacenen en ningún lado, sigue siendo un tratamiento de datos personales. Para poder trabajar con esto, amparados por una base legal, nos podríamos ceñir a lo siguiente:
   1. Consentimiento: La persona que llama debe ser informada de que su voz va a ser analizada y con qué propósito. Algo así como: “le informamos que esta llamada puede ser grabada con el fin de analizar la voz por motivos de seguridad y prevención de fraude. Si continua en la llamada acepta el análisis”.
   2. Interés legítimo: Frente a un juez, se tendría que justificar “El Juicio de Ponderación”, donde se pesa el interés contra los derechos de las personas que nos llaman, teniendo así que justificar:
      1. ¿Es necesario?
      2. ¿Es proporcionado?
      3. ¿Prevalece de interés?

Es decir, se actuaría sin permiso directo. Existe el riesgo de que la base jurídica del tratamiento sea cuestionada, y es que, si una autoridad considera en el balance de intereses que el derecho de protección de datos del usuario prevalece sobre el interés de seguridad, no se justificaría por este método, y para cumplir con las normativas vigentes, se debería actuar por el método 2.a.